# A novel feature selection score for text categorization

**Susana Eyheramendy**
Department of Statistics
1 South Parks Road
Oxford University
Oxford, OX1 3TG

**David Madigan**
Department of Statistics
501 Hill Center
Rutgers University
Piscataway, NJ 08855

## Abstract

This paper proposes a new feature selection score for text classification. The value that this score assigns to each feature has an appealing Bayesian interpretation, being the posterior probability of inclusion of the feature in a model. We evaluate the performance of the score, together with five other feature selection scores that have been prominent in the text categorization literature, using four classification algorithms and two benchmark text datasets. We find that the new score performs well although no one score dominates all others.

## 1 Introduction

The text classification literature tends to focus on feature selection algorithms that compute a score independently for each candidate feature. This is the so-called *filtering* approach. The scores typically contrast the counts of occurrences of words or other linguistic artifacts in training documents that belong to the target class with the same counts for documents that do not belong to the target class. Given a predefined number of words to be selected, say $d$, one chooses the $d$ words with the highest score. Several score functions exist (Section 3 provides definitions). Yang and Pedersen (1997) show that Information Gain and $\chi^2$ statistics performed best among five different scores. Forman (2003) provides evidence that these two scores have correlated failures. Hence when choosing optimal pairs of scores these two scores work poorly together. He introduced a new score, the Bi-Normal Separation, that yields the best performance on the greatest number of tasks among twelve feature selection scores. Mladenic and Grobelnik (1999) compare eleven scores combined using a Naive Bayes classifier and find that the Odds Ratio score performed best in the highest number of tasks.

In regression and classification problems in Statistics, popular feature selection strategies depend on the same algorithm that fits the models. This is the so-called *wrapper* approach. For example, *Best subset regression* finds for each $k$ the best subset of size $k$ based on residual sum of squares. *Leaps and bounds* is an efficient algorithm that finds the best set of features when the number of predictors is no larger than about 40. Miller (2002) provides an extensive discussion.

Barbieri and Berger (2004) in a Bayesian context and under certain assumptions show that for selection among normal linear models, the best model contains those features which have overall posterior probability greater than or equal to 1/2. Motivated by this study we introduce a new feature selection score (PIP) that evaluates the posterior probability of inclusion of a given feature over all possible models, where the models correspond to a set of features. Unlike typical scores used for feature selection via filtering, the PIP score does depend on a specific model. In this sense, the new score straddles the filtering and wrapper approaches.

We present experiments that compare the new feature selection score with five other feature selection scores that have been prominent in the studies mentioned above. We evaluate these feature selection scores on two widely-used benchmark text classification datasets, Reuters-21578 and 20-Newsgroups, with four classification algorithms. Following previous studies, we measure the performance of the classification algorithms using the $F_1$ measure.

We have organized this paper as follows. Section 2 briefly presents the theory that motivates the new feature selection score. Section 3 describes the various feature selection scores we consider, both the new score and the various existing competitors. In Section 4 we mention the classification algorithms that we use to compare the feature selection scores. The experimental settings and experimental results are in Section 5. Section 6 has the conclusions.

## 2 Motivation for the new feature selection score (PIP)

In this section we present the theory behind the median probability model introduced by Barbieri and Berger (2004) that motivates our work. Consider the usual normal linear model:

$$y = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $y$ is the $n \times 1$ vector of observed values of the response variable, $X$ is the $n \times k$ $(k < n)$ full rank design matrix of covariates, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown coefficients. Assume that the coordinates of the random error vector are independent, each with a normal distribution with mean 0 and variance $\sigma^2$.

We call the model in equation (1) the *full* model and consider selecting a model from among all submodels of the form $M_l : y = X_l\boldsymbol{\beta}_l + \boldsymbol{\epsilon}$, where $l = (l_1, ..., l_k)$ is the model index, $l_i$ being either 1 or 0 as covariate $x_i$ is in or out of the model; $X_l$ contains the columns of $X$ corresponding to the nonzero coordinates of $l$; and $\boldsymbol{\beta}_l$ is the corresponding vector of regression coefficients.

For a future vector of covariates $x^* = (x_1^*, ..., x_k^*)$, we assume that the loss in predicting $y^* = x^*\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by $\hat{y}^*$ is the squared error loss $L(\hat{y}^*, y^*) = (\hat{y}^* - y^*)^2$.

Assume also that covariates $x^*$ arise according to some distribution and that the $k \times k$ matrix:

$$Q = E(x^{*T}x^*), \tag{2}$$

exists and is positive definite.

The optimal predictor of $y^*$, under squared error loss and when the model $M_l$ is true, is given by

$$\hat{y}_l^* = x^* H_l \tilde{\boldsymbol{\beta}}_l, \tag{3}$$

where $\tilde{\boldsymbol{\beta}}_l$ is the posterior mean of $\boldsymbol{\beta}_l$ with respect to $\pi_l(\boldsymbol{\beta}_l, \sigma|y)$, the posterior distribution of the unknown parameters in $M_l$. $H_l$ is the matrix such that $xH_l$ is the subvector of $x$ corresponding to the nonzero coordinates of $l$, i.e., the covariate vector corresponding to model $M_l$.

When one must select a single model, under the Bayesian approach, a common perception exists that the optimal predictive model is the model with the hightest posterior probability. However, this is not necessarily the case. For selection among normal linear models, the optimal predictive model is often the median probability model, which we define in what follows.

**Definition 1** The posterior inclusion probability for variable $x_i$ is

$$p_i = \sum_{l:l_i=1} P(M_l|y)$$

**Definition 2** If it exists, the median probability model, $M_{l^*}$, is the model that contains all those variables whose posterior inclusion probability is at least 1/2. More precisely, $l^*$ is such that

$$l_i^* = \begin{cases} 1 & \text{if } p_i \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

**Theorem** Suppose we select a single model to predict a future observation. If:

*i)* $Q$ (as in (2)) is diagonal with diagonal elements $q_i > 0$, and
*ii)* $\tilde{\boldsymbol{\beta}}_l = H_l\tilde{\boldsymbol{\beta}}$ where $\tilde{\boldsymbol{\beta}}_l$ is defined in (3) (i.e. that the posterior mean of $\boldsymbol{\beta}_l$ correspond to the relevant coordinates of the posterior mean in the full model),

then the median probability model is the best predictive model.

The results in Barbieri and Berger (2004) do not directly apply to the models that we consider. We do not consider normal linear models and furthermore $Q$ is rarely a diagonal matrix in practice. Nonetheless, the remarkable results in Barbieri and Berger (2004) do suggest that the median probability model certainly warrants consideration even in situations where the conditions do not strictly apply. In what follows we present a novel algorithm for computing the posterior inclusion probability for different text categorization models.

## 3 Feature Selection Scores

Feature selection, or word selection in the experiments of this study, uses a score to select the best $d$ words from all words that appear in the training set. Before we list the feature selection scores that we study, we introduce some notation. Table 1 show the basic statistics for a single word and a single category (or class).

$n_{kw}$ : $n^\circ$ of documents in class $c_k$ with word $w$.
$n_{k\overline{w}}$ : $n^\circ$ of documents in class $c_k$ without word $w$.
$n_{\overline{k}w}$ : $n^\circ$ of documents not in class $c_k$ with word $w$.
$n_{\overline{k}\overline{w}}$ : $n^\circ$ of documents not in class $c_k$ without word $w$.

|  | $c_k$ | $c_{\overline{k}}$ |  |
|---|---|---|---|
| $w$ | $n_{kw}$ | $n_{\overline{k}w}$ | $n_w$ |
| $\overline{w}$ | $n_{k\overline{w}}$ | $n_{\overline{k}\overline{w}}$ | $n_{\overline{w}}$ |
|  | $n_k$ | $n_{\overline{k}}$ | $n$ |

Table 1: Two-way contingency table of word $w$ and category $c_k$

$n_k$ : total $n^\circ$ of documents in class $c_k$.
$n_{\overline{k}}$ : total $n^\circ$ of documents that are not in class $c_k$.
$n_w$ : total $n^\circ$ of documents with word $w$.
$n_{\overline{w}}$ : total $n^\circ$ of documents without word $w$.
$n$ : total $n^\circ$ of documents.

### 3.1 Posterior Inclusion Probability (PIP) under a Bernoulli distribution

We introduce a new feature selection score which is motivated by the median probability model. We first consider the binary naive Bayes model. Section 3.2 considers a naive Bayes model with Poisson distributions for word frequency. This score for feature or word $w$ and class $c_k$ is defined as

$$PIP(w, c_k) = \frac{l_{0wk}}{l_{0wk} + l_{wk}} \tag{4}$$

where

$$\begin{aligned} l_{0wk} &= \frac{B(n_{kw} + \alpha_{kw}, n_{k\overline{w}}\beta_{kw})}{B(\alpha_{kw}, \beta_{kw})} \\ &\quad \times \frac{B(n_{\overline{k}w} + \alpha_{\overline{k}w}, n_{\overline{k}\overline{w}} + \beta_{\overline{k}w})}{B(\alpha_{\overline{k}w}, \beta_{\overline{k}w})} \\ l_{wk} &= \frac{B(n_w + \alpha_w, n_{\overline{w}} + \beta_w)}{B(\alpha_w, \beta_w)} \end{aligned}$$

$B(a,b)$ is the *Beta* function which is defined as $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, and $\alpha_{kw}$, $\alpha_{k\overline{w}}$, $\alpha_w$, $\beta_{kw}$, $\beta_{k\overline{w}}$, $\beta_w$ are constants set by the practitioner. In our experiments we set them to be $\alpha_w = 0.2$, $\beta_w = 2/25$ for all words $w$, $\alpha_{kw} = 0.1$, $\alpha_{k\overline{w}} = 0.1$, $\beta_{kw} = 1/25$ and $\beta_{k\overline{w}} = 1/25$ for all categories $k$ and words $w$. These settings correspond to rather diffuse priors.

We explicate this score on the context of a two-candidate-word model. In general, with $d$ candidate words, there are $2^d$ models corresponding to allpossible subsets of the words. For two words, Figure 1 we show a graphical representation of the four possible models. The corresponding likelihoods for each model are given by

$M_{(1,1)} : \prod_i Pr(w_{i1}, w_{i2}, c_i|\theta_{1c}, \theta_{2c}) = \prod_i \mathcal{B}(w_{i1}, \theta_{k1})$
$\times\mathcal{B}(w_{i1}, \theta_{\overline{k}1})\mathcal{B}(w_{i2}, \theta_{k2})\mathcal{B}(w_{i2}, \theta_{\overline{k}2})Pr(c_i|\theta_k)$

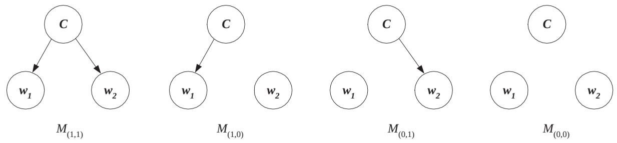

Figure 1: Graphical model representation of the four models with two words, $w_1$ and $w_2$.

$M_{(1,0)} : \prod_i Pr(w_{i1}, w_{i2}, c_i|\theta_{1c}, \theta_2) = \prod_i \mathcal{B}(w_{i1}, \theta_{k1})$
$\times\mathcal{B}(w_{i1}, \theta_{\overline{k}1})\mathcal{B}(w_{i2}, \theta_2)\mathcal{B}(w_{i2}, \theta_2)Pr(c_i|\theta_k)$
$M_{(0,1)} : \prod_i Pr(w_{i1}, w_{i2}, c_i|\theta_1, \theta_{2c}) = \prod_i \mathcal{B}(w_{i1}, \theta_1)$
$\times\mathcal{B}(w_{i1}, \theta_1)\mathcal{B}(w_{i2}, \theta_{k2})\mathcal{B}(w_{i2}, \theta_{\overline{k}2})Pr(c_i|\theta_k)$
$M_{(0,0)} : \prod_i Pr(w_{i1}, w_{i2}, c_i|\theta_1, \theta_2) = \prod_i \mathcal{B}(w_{i1}, \theta_1)$
$\times\mathcal{B}(w_{i1}, \theta_1)\mathcal{B}(w_{i2}, \theta_2)\mathcal{B}(w_{i2}, \theta_2)Pr(c_i|\theta_k)$

where $w_{ij}$ takes the value 1 if document $i$ contains word $j$ and 0 otherwise, $c_i$ is 1 if document $i$ is in category $k$ otherwise is 0, $Pr(c_i|\theta_k) = \mathcal{B}(c_i, \theta_k)$ and $\mathcal{B}(w, \theta) = \theta^w(1-\theta)^{1-w}$ denotes a Bernoulli probability distribution.

Therefore, in model $M_{(1,1)}$ the presence or absence of both words in a given docuement depends on the document class. $\theta_{k1}$ corresponds to the proportion of documents in category $c_k$ with word $w_1$ and $\theta_{\overline{k}1}$ to the proportion of documents not in category $c_k$ with word $w_1$. In model $M_{(1,0)}$ only word $w_1$ depends on the category of the document and $\theta_2$ correspond to the proportion of documents with word $w_2$ regardless of the category associated with them. $\theta_k$ is the proportion of documents in category $c_k$ and $Pr(c_i|\theta_k)$ is the probability that document $d_i$ is in category $c_k$.

We assume the following prior probability distributions for the parameters,

$\theta_{kw} \sim Beta(\alpha_{kw}, \beta_{kw})$
$\theta_{\overline{k}w} \sim Beta(\alpha_{\overline{k}w}, \beta_{\overline{k}w})$
$\theta_w \sim Beta(\alpha_w, \beta_w)$
$\theta_k \sim Beta(\alpha_k, \beta_k)$

where $Beta(\alpha, \beta)$ denotes a Beta distribution, i.e. $Pr(\theta|\alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, $k \in \{1, ..., m\}$ and $w \in \{1, ..., d\}$.

Then the marginal likelihoods for each of the four models above are:

$Pr(data|M_{(1,1)}) = l_0 \times l_{01k} \times l_{02k}$
$Pr(data|M_{(1,0)}) = l_0 \times l_{01k} \times l_{2k}$
$Pr(data|M_{(0,1)}) = l_0 \times l_{1k} \times l_{02k}$
$Pr(data|M_{(0,0)}) = l_0 \times l_{1k} \times l_{2k}$

where $l_{0wk}$ and $l_{wk}$ are defined above for $w \in \{1, 2, ..., d\}$ and $l_0 = \int_0^1 \prod_i Pr(c_i|\theta_k)Pr(\theta_k|\alpha_k, \beta_k)d\theta_k$ is the marginal probability for the category of the documents.

The overall posterior probability that a feature is included in a model, its posterior inclusion probability (PIP), is defined as

$$PIP(w, c_k) = \sum_{l:l_j=1} Pr(M_l|data) \qquad (5)$$

where $l$ is a vector of length the number of features and the $j$th component takes the value 1 if the $j$th feature is included in model $M_l$, otherwise it is 0. It is straightforward to show that $PIP(w, c_k)$ in equation (4) is equivalent to $PIP(w, c_k)$ in equation (5), if we assume that the prior probability density for the models is uniform, e.g. $Pr(M_l) \propto 1$.

In the example above, the posterior inclusion probability for word $w_1$ is given by,

$$\begin{aligned} Pr(w_1|c_k) &= Pr(M_{(1,1)}|data) + Pr(M_{(1,0)}|data) \\ &= \frac{l_{01k}}{l_{01k} + l_{1k}} \end{aligned}$$

To get a single "bag of words" for all categories we compute the weighted average of $PIP(w, c_k)$ over all categories.

$$PIP(w) = \sum_k Pr(c_k) PIP(w, c_k)$$

We note that Dash and Cooper (2002) present similar manipulations of the naive Bayes model but for model averaging purposes rather than finding the median probability model.

### 3.2 Posterior Inclusion Probability (PIPp) under Poisson distributions

A gernalization of the binary naive Bayes model assumes class-conditional Poisson distributions for the word frequencies in a document. As before, assume that the probability distribution for a word in a document might or might not depend on the category of the document. More precisely, if the distribution for word $w$ depends on the category $c_k$ of the document we have,

$$\begin{aligned} Pr(w|c=1) &= \frac{e^{-\lambda_{kw}} \lambda_{kw}^{w}}{w!} \\ Pr(w|c=0) &= \frac{e^{-\lambda_{\overline{k}w}} \lambda_{\overline{k}w}^{w}}{w!} \end{aligned}$$

where $w$ denotes a specific word and the number of times that word appears in the document and $\lambda_{kw}$ ($\lambda_{\overline{k}w}$) represents the expected number of times that word $w$ appears in documents in category $c_k$ ($c_{\overline{k}}$). If the distribution for word $w$ does not depend on the category of the document then we have,

$$Pr(w) = \frac{e^{-\lambda_w} \lambda_w^{w}}{w!}$$

where $\lambda_w$ represents the expected number of times $w$ appears in a document regardless of the category of the document.

Assume the following conjugate prior probability densities for the parameters,

$$\begin{aligned} \lambda_{kw} &\sim Gamma(\alpha_{kw}, \beta_{kw}) \\ \lambda_{\overline{k}w} &\sim Gamma(\alpha_{\overline{k}w}, \beta_{\overline{k}w}) \\ \lambda_w &\sim Gamma(\alpha_w, \beta_w) \end{aligned}$$

where $\alpha_{kw}, \beta_{kw}, \alpha_{\overline{k}w}, \beta_{\overline{k}w}, \alpha_w, \beta_w$ are hyperparameters to be set by the practitioner.

Now, as before, the posterior inclusion probability for poisson distributions (PIPp) is given by

$$PIPp(w, c_k) = \frac{l_{0wk}}{l_{0wk} + l_{wk}}$$

where

$$\begin{aligned} l_{0wk} &= \frac{\Gamma(N_{kw} + \alpha_{kw})}{\Gamma(\alpha_{kw})\beta_{kw}^{\alpha_{kw}}} \frac{\Gamma(N_{\overline{k}w} + \alpha_{\overline{k}w})}{\Gamma(\alpha_{\overline{k}w})\beta_{\overline{k}w}^{\alpha_{\overline{k}w}}} \\ &\quad \times (\frac{\beta_{kw}}{n_k \beta_{kw} + 1})^{n_{kw} + \alpha_{kw}} (\frac{\beta_{\overline{k}w}}{n_{\overline{k}} \beta_{\overline{k}w} + 1})^{n_{\overline{k}w} + \alpha_{\overline{k}w}} \\ l_{wk} &= \frac{\Gamma(N_w + \alpha_w)}{\Gamma(\alpha_w)} (\frac{\beta_w}{\beta_w n + 1})^{n_w + \alpha_w} \frac{1}{\beta_w^{\alpha_w}} \end{aligned}$$

This time, $N_{kw}, N_{\overline{k}w}, N_w$ denote:

$N_{kw}$: $n^\circ$ of times word $w$ appears in documents in class $c_k$.
$N_{\overline{k}w}$: $n^\circ$ of times word $w$ appears in documents not in class $c_k$.
$N_w$: total $n^\circ$ of times that word $w$ appears in all documents.

As before, to get a single "bag of words" for all categories we compute the weighted average of $PIPp(w, c_k)$ over all categories.

$$PIPp(w) = \sum_k Pr(c_k) PIPp(w, c_k)$$

### 3.3 Information Gain (IG)

Information gain is a popular score for feature selection in the field of machine learning. In particular it is used in the C4.5 decision tree inductive algorithm.

Yang and Pedersen (1997) compare five different feature selection scores on 2 datasets and show that Information Gain is among the two most effective ones. The information gain of word $w$ is defined to be:

$$\begin{aligned} IG(w) &= -\sum_{k=1}^{m} Pr(c_k) \log Pr(c_k) \\ &\quad + Pr(w) \sum_{k=1}^{m} Pr(c_k|w) \log Pr(c_k|w) \\ &\quad + Pr(\overline{w}) \sum_{k=1}^{m} Pr(c_k|\overline{w}) \log Pr(c_k|\overline{w}) \end{aligned}$$

where $\{c_k\}_{k=1}^{m}$ denote the set of categories and $\overline{w}$ the abscence of word $w$. It measures the decrease in entropy when the feature is present versus when the feature is absent.

### 3.4 Bi-Normal Separation (BNS)

Forman (2003) defines Bi-Normal Separation as:

$$BNS(w, c_k) = |\Phi^{-1}(\frac{n_{kw}}{n_k}) - \Phi^{-1}(\frac{n_{\overline{k}w}}{n_{\overline{k}}})|$$

where $\Phi$ is the standard normal distribution and $\Phi^{-1}$ its corresponding inverse. $\Phi^{-1}(0)$ is set to be equal to 0.0005 to avoid numerical problems following Forman (2003). By averaging over all categories, we get a score that selects a single set of words for all categories.

$$BNS(w) = \sum_{k=1}^{m} Pr(c_k) |\Phi^{-1}(\frac{n_{kw}}{n_k}) - \Phi^{-1}(\frac{n_{\overline{k}w}}{n_{\overline{k}}})|$$

### 3.5 Chi-Square

The chi-square feature selection score, $\chi^2(w, c_k)$, measures the dependence between word $w$ and category $c_k$. If word $w$ and category $c_k$ are independent $\chi^2(w, c_k)$ is equal to zero. When we select a different set of words for each category we utilise the following score,

$$\chi^2(w, c_k) = \frac{n(n_{kw} n_{\overline{k}\overline{w}} - n_{\overline{k}w} n_{k\overline{w}})^2}{n_k n_w n_{\overline{k}} n_{\overline{w}}}.$$

Again, by averaging over all categories we get a score for selecting a single set of words for all categories.

$$\chi^2(w) = \sum_{k=1}^{m} Pr(c_k) \chi^2(w, c_k).$$

### 3.6 Odds Ratio

The Odds Ratio measures the odds of word $w$ occuring in documents in category $c_k$ divided by the odds of word $w$ not occuring in documents in category $c_k$. Mladenic and Grobelnik (1999) find this to be the best score among eleven scores for a Naive Bayes classifier. For category $c_k$ and word $w$ the Odds Ratio (OR) is given by,

$$OR(w, c_k) = \frac{\frac{n_{kw}+0.1}{n_k+0.1} / \frac{n_{k\overline{w}}+0.1}{n_k+0.1}}{\frac{n_{\overline{k}w}+0.1}{n_{\overline{k}}+0.1} / \frac{n_{\overline{k}\overline{w}}+0.1}{n_{\overline{k}}+0.1}}$$

where we added the constant 0.1 to avoid numerical problems. By averaging over all categories we get,

$$OR(w) = \sum_{k} Pr(c_k) OddsRatio(w, c_k).$$

### 3.7 Word Frequency

This is the simplest of the feature selection scores. In the study of Yang and Pedersen (1997) they show that word frequency is the third best after information gain and $\chi^2$. They also point out that there is strong correlation between these two scores and word frequency. For each category $c_k$ word frequency for word $w$, is the number of documents in $c_k$ that contain word $w$, i.e. $WF(w, c_k) = n_{kw}$.

Averaging over all categories we get a score for each $w$,

$$WF(w) = \sum_{k} Pr(c_k) WF(w, c_k) = \sum_{k} Pr(c_k) n_{kw}.$$

## 4 Classification Algorithms

To determine the performance of the different feature selection scores, the classification algorithms that we consider are the Multinomial, Poisson and Binary Naive Bayes classifiers ( McCallum and Nigam, 1998, Lewis, 1998, and Eyheramendy *et al*, 2003) and the hierarchical probit classifier of Genkin et al (2003). The naive Bayes models are generative models (i.e., models for $Pr(x, y)$) while the probit is a discriminative model (i.e., a model for $Pr(y|x)$). Many text classification applications continue to utilize Naive Bayes models. However, discriminative models such as support vector machines and the hierarchical probit classifer typically provide superior predictive performance. Genkin *et al.* (2003) provide detailed experimental results.

## 5 Experimental Settings and Results

Before we start the analysis we remove common non-informative words taken from a standard *stopword* list of 571 words and we remove words that appear less

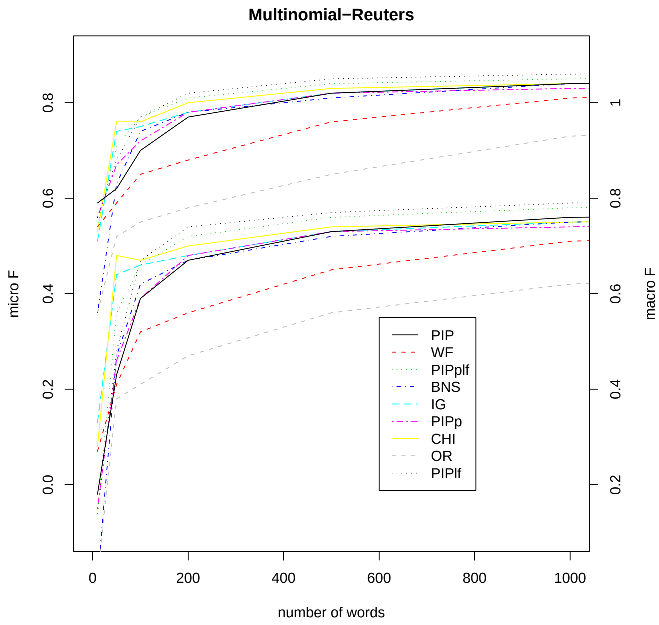

Figure 2: Curves of performance for the multinomial model for different number of words measure by macro $F_1$ and micro $F_1$ (which correspond to the bottom and top set of curves resp.).

than three times in the training documents, justifying this with the fact that they are unlikely to appear in testing documents. This eliminates $8,752$ words in the Reuters dataset (38% of all words in training documents) and $47,118$ words in the Newsgroups dataset (29% of all words in training documents). Words appear on average in 1.41 documents in the Reuters dataset and in 1.55 documents in the Newsgroups dataset.

### 5.1 Datasets

The 20-Newsgroups dataset contains $19,997$ articles divided almost evenly into 20 disjoint categories. The categories topics are related to computers, politics, religion, sport and science. We split the dataset randomly into 75% for training and 25% for testing. We took this version of the dataset from http://www.ai.mit.edu/people/jrennie/20Newsgroups/. The other dataset comprises a subset of the ModApte version of the Reuters$-21,578$ collection, where each document has assigned at least one topic label (or category) and this topic label belongs to any of the 10 most populous categories - earn, acq, grain, wheat, crude, trade, interest, corn, ship, money-fx. It contains $6,775$ documents in the training set and $2,258$ in the testing set.

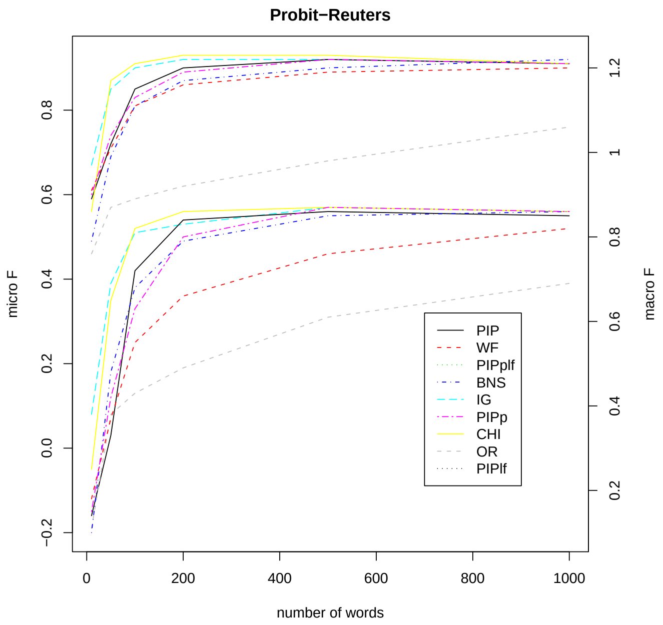

Figure 3: Curves of performance for the probit model for different number of words measure by macro and micro $F_1$ (top and bottom sets of curves resp.) for the Reuters dataset.

### 5.2 Experimental Results

In these experiments we compare seven feature selection scores, on two benchmark datasets, Reuters-21578 and Newgroups (see subsection 5.1), under four classification algorithms (see section 4). We report so-called $F_1$ performance measures. $F_1$ is the average of precision and recall. See, for example, Genkin *et al.* (2003) for details.

We compare the performance of the classifiers for different numbers of words and vary the number of words from 10 to 1000. For larger number of words the classifiers tend to perform somewhat more similarly, and the effect of chosing the words using a different feature selection procedure is less noticeable.

Figure 2, 3, 4 and 5 show the micro and macro averaged $F_1$ measure for each of the feature selection scores as we vary the number of features to select for the four classification algorithms - multinomial, probit, poisson and binary respectively. In order to have both sets of curves (the curves with the micro $F_1$ and macro $F_1$ measures) in the same graph we move them apart. The $y-axes$ for the micro $F_1$ (macro $F_1$) measure correspond to the $y-axes$ on the left (right). The reader will find these figures easier to read in a color rather than black and white rendition.

We noticed that PIP gives, in general, high values to all very frequent words. This lead us to consider a second version of PIP and PIPp, PIPlf and PIPplf

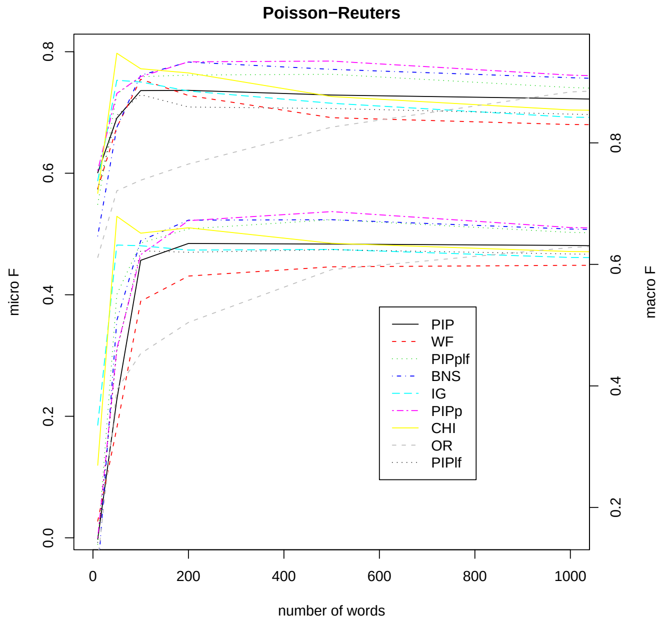

Figure 4: Curves of performance for the poisson model for different number of words measure by micro $F_1$ and macro $F_1$ (top and bottom sets of curves resp.) for the Reuters dataset.

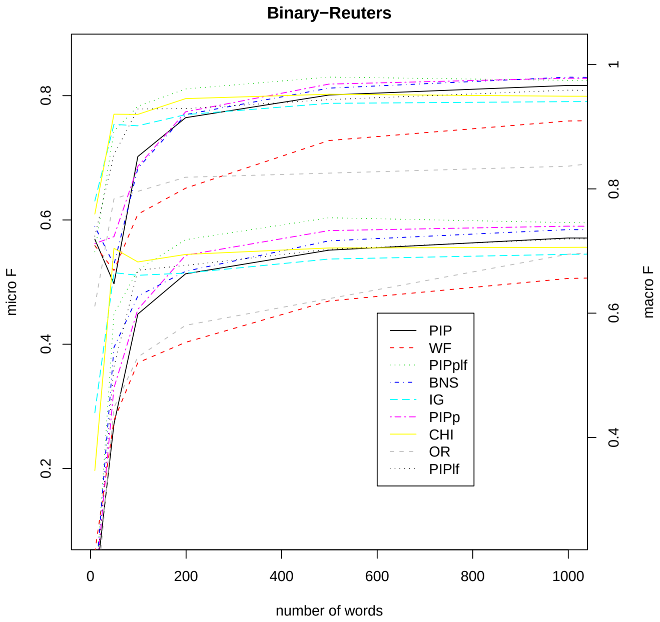

Figure 5: Curves of performance for the binary naive Bayes model for different number of words measure by micro $F_1$ and macro $F_1$ (top and bottom sets of curves resp.) for the Reuters dataset.

respectively, which correspond to the same score but with the words that appear too frequently removed. Specifically, we remove words that appear more than 2000 times in the Reuters dataset (that accounts for 15 words) and more than 3000 times in the Newsgroups dataset (that accounts for 36 words).

**Reuters**. Like the results of Forman (2003), if for scalability reasons one is limited to a small number of features ($< 50$) the best available metrics are IG and $\chi^2$ as Figures $2 - 5$ show. For larger number of features ($> 50$), Figure 2 shows that PIPplf and PIPlf are the best scores for the mutinomial classifier. Figure 4 and 5 show the performance for the poisson and binary classifiers respectively. PIPp followed by BNS achive the best performance in the Poisson classifier and PIPplf achieves the best performance in the binary classifier. WF performs poorly compare to the other scores in all the classifiers, having the best performance with the poisson.

**Newsgroups**. $\chi^2$ followed by BNS, IG and PIP are the best performing measures in the probit classifier. $\chi^2$ is also the best one in multinomial model followed by BNS and in the binary classifier with the macro $F_1$ measure. OR performs best in the poisson classifier. PIPp is best in the binary classifier under the micro $F_1$ measure. WF performs poorly compare to the other scores in all classifiers. Because of lack of space we do not show graphical display of the performance of the classifiers in the Newsgroups dataset.

Table 2 shows the performance of the four classifiers in the two datasets with 1,000 features. For the Reuters dataset, BNS provides the best performance for three of models. However, a PIP score ties BNS in two cases, and comes close in a third. PIPlf provides the best performance in one case. For the newsgroup data, OR and BNS are best.

Table 3 shows the predictive performance with 200 features. Again, PIP scores do well on Reuters and less well on Newsgroups. The chi-square score provides the best performance in two cases.

## 6 Conclusion

In this study we introduced a new feature selection score, PIP. The value that this score assigns to each word has an appealing Bayesian interpretation, being the posterior probability of inclusion of the word in a model. Such models assume a probability distribution on the words of the documents. We consider two probability distributions, Bernoulli and Poisson. The former takes into account the presence or absence of words in the documents, and the latter, the number of times each word appears in the documents. Future research could consider alternative PIP socres corresponding to different probabilistic models.

$\chi^2$, BNS, and PIP are the best performing scores. Still, feature selection scores and classification algo-

| | IG | PIP | $\chi^2$ | OR | BNS | WF | PIP$_p$ | PIP$_{lf}$ | PIP$_{plf}$ |
|---|---|---|---|---|---|---|---|---|---|
| | Reuters-21578 dataset | | | | | | | | |
| probit | 0.91 | 0.91 | 0.91 | 0.76 | **0.92** | 0.90 | 0.91 | 0.90 | 0.91 |
| poisNB | 0.69 | 0.72 | 0.70 | 0.73 | **0.76** | 0.68 | **0.76** | 0.70 | 0.74 |
| multiNB | 0.84 | 0.84 | 0.84 | 0.73 | 0.84 | 0.81 | 0.83 | **0.86** | 0.85 |
| binNB | 0.79 | 0.82 | 0.80 | 0.69 | **0.83** | 0.76 | **0.83** | 0.81 | 0.82 |
| | 20-Newsgroup dataset | | | | | | | | |
| probit | 0.75 | 0.74 | 0.77 | 0.63 | **0.76** | 0.64 | 0.72 | 0.74 | 0.72 |
| poisNB | 0.77 | 0.77 | 0.81 | **0.93** | 0.80 | 0.68 | 0.85 | 0.77 | 0.85 |
| multiNB | 0.58 | 0.59 | 0.62 | 0.50 | **0.64** | 0.61 | 0.55 | 0.62 | 0.58 |
| binNB | 0.55 | 0.52 | 0.58 | **0.59** | 0.56 | 0.46 | 0.57 | 0.52 | 0.57 |

Table 2: The first column correspond to the classifier (probit,poisson,multinomial,binary. The numbers on the other columns correspond to the micro $F_1$ measure for 1000 words.

| | IG | PIP | $\chi^2$ | OR | BNS | WF | PIP$_p$ | PIP$_{lf}$ | PIP$_{plf}$ |
|---|---|---|---|---|---|---|---|---|---|
| | Reuters-21578 dataset | | | | | | | | |
| probit | 0.92 | 0.90 | **0.93** | 0.62 | 0.87 | 0.86 | 0.89 | 0.90 | 0.87 |
| poisNB | 0.74 | 0.74 | 0.77 | 0.62 | **0.78** | 0.73 | **0.78** | 0.71 | 0.76 |
| multiNB | 0.78 | 0.77 | 0.80 | 0.58 | 0.78 | 0.68 | 0.78 | **0.82** | 0.81 |
| binNB | 0.77 | 0.76 | 0.80 | 0.67 | 0.77 | 0.65 | 0.77 | 0.78 | **0.81** |
| | 20-Newsgroup dataset | | | | | | | | |
| probit | 0.66 | 0.65 | **0.70** | 0.53 | 0.66 | 0.46 | 0.59 | 0.59 | 0.53 |
| poisNB | 0.80 | 0.82 | 0.84 | **0.94** | 0.86 | 0.38 | 0.84 | 0.83 | 0.84 |
| multiNB | 0.52 | 0.53 | **0.58** | 0.51 | 0.55 | 0.37 | 0.53 | 0.53 | 0.47 |
| binNB | 0.53 | 0.50 | 0.56 | 0.43 | **0.58** | 0.36 | **0.58** | 0.53 | **0.58** |

Table 3: The first column correspond to the classifier (probit,poisson,multinomial,binary. The numbers on the other columns correspond to the micro $F_1$ measure for 200 words.

rithms seem to be highly data- and model-dependent. The feature selection literature reports similarly mixed findings. For instance, Yang and Pedersen (1997) find that IG and $\chi^2$ are the strongest feature selection scores. They perform their experiments on two datasets, Reuters-22173 and OHSUMED, and under two classifiers, kNN and a linear least square fit. Mladenic and Grobelnik (1999) find that OR is the strongest feature selection score. They perform their experiments on a Naive Bayes model and use the Yahoo dataset. Forman (2003) favors bi-normal separation.

Our results regarding the performance of the different scores are consistent with Yang and Pedersen (1997) in that $\chi^2$ and IG seem to be strong scores for feature selection in discriminative models, but disagree in that WF appears to be a weak score in most instances. Note that we do not use exactly the same WF score. Ours is a weighted average by the category proportion.

## Acknowledgements

We are grateful to David D. Lewis for helpful discussions.

## References

Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *Annals of Statistics*, **32**, 870–897.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Dash, D. and Cooper, G.F. (2002). Exact model averaging with naive Bayesian classifiers. In: *Proceedings of the Nineteenth International Conference on Machine Learning*, 91-98.

Eyheramendy, S., Lewis, D.D. and Madigan, D. (2003). On the naive Bayes classifiers for text categorization. In *Proceedings of the ninth international workshop on Artificial Intelligence and Statistics*, eds, C.M. Bishop and B.J. Frey.

Forman, G. (2003). An extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*

Genkin, A., Lewis, D.D., and Madigan, D. (2003). Large-scale Bayesian Logistic Regression for Text Classification. `stat.rutgers.edu/~madigan/PAPERS/shortFat-v13.ps`

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of ECML-98, 137–142.

Lewis, D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. Proceedings of ECML-98, 4–15.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI/ICML Workshop on Learning for Text Categorization*, pages 41 – 48.

Miller, A.J. (2002) *Subset selection in regression (second edition)*. Chapman and Hall.

Mladenic, D. and Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive Bayes. Proceedings ICML-99, pages 258-267.

Silvey, S. D. (1975). Statistical Inference. Chapman & Hall. London.

Yang, Y. and Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. Proceedings ICML-97, 412-420.